SYSTEM, METHOD, AND PROGRAM FOR ESTIMATING GENE EXPRESSION STATE, AND RECORDING MEDIUM THEREFOR

Background of the Invention:

[0001] The present invention relates to a method and system for statistical analysis of cDNA microarray data using two different fluorescence dyes, and a recording medium for the same. In particular, the present invention relates to a system, method, and program for estimating the probability of gene expression in each channel, and a recording medium for the same.

[0002] Currently, the study of genomics is expanding from structural analysis on individual genes to systematic functional analysis of genes. Experiments using cDNA (complementary DNA) microarrays capable of simultaneously quantifying the expression levels of a large number of genes are expected to be extremely effective in functional analysis of functionally unknown genes or whole genes.

[0003] The objective of experiments using cDNA microarrays with two different fluorescence dyes is to detect the difference in gene expression level between two kinds of cells. The following gives a summary of a cDNA microarray configuration with two different fluorescence dyes. First, cDNAs of a large number of sets of genes are densely fixed on glass slides in arrays (microarrays) as reference probes.

[0004] Next, mRNAs extracted from two kinds of different conditional samples, cell 1 and cell 2 (e.g., normal cell and cancer cell), are labeled respectively with fluorescence dyes different in wavelength from each other to synthesize target cDNAs. Then, these cDANs are mixed in equal proportions,

and hybridized with the microarrayed cDNAs or reference probes. After this competitive hybridization, the glass slides are imaged using a scanner and fluorescence intensities are measured separately for each dye. The fluorescence dye with which the cell 1 is labeled and the fluorescence dye with which the cell 2 is labeled are read from channel 1 and channel 2, respectively, to obtain gene expression level data (microarray data).

[0005] Thus, since the process of obtaining microarray data is so complicated as to require advanced experimental techniques, it is conceivable that several experimental errors could occur at each stage of the experiment. Therefore, in order to retrieve data to be truly biologically significant from the microarray data, analyzing expression level distributions and experimental errors presents a significant challenge to be solved.

[0006] In regard to the expression level distributions, prior art document 1 (Newton et. al., 2001, *Journal of Computational Biology*, Vol. 8, pp. 37-52) can be referred to, for example, in which Newton et. al. forms a hypothesis that proposes the use of the gamma distribution function to help analyze expression levels so as to consider statistical characteristics about the ratio of expression levels (the ratio of expression levels in channel 1 and channel 2).

$$f(x) = p\varphi\left(x - \mu_1 \mid \sigma_1^2\right) + (1 - p)\varphi\left(x - \mu_2 \mid \sigma_2^2\right) \quad (1)$$

[0007] In regard to the observed expression level data, prior art document 2 (Lee et. al., 2000, *Proceeding of the National Academy of Sciences*, Vol. 97, No. 18, pp. 9834-9839) can be referred to, for example. Assuming the ability to separate true expression levels into two levels and the existence of accidental errors, Lee et. al. adopts a mixed normal distribution as shown in the following equation (1) to consider statistical characteristics about the expression level data:

[0008] Here, x denotes (the logarithmic value of) a gene expression level such as fluorescence intensity obtained with a scanner or the like.

Further, the first term of the right hand side, $\varphi\left(x - \mu_1 \mid \sigma_1^2\right)$, represents a normal distribution with average μ_1 and variance σ_1^2 when a gene is being expressed, the second term $\varphi\left(x - \mu_2 \mid \sigma_2^2\right)$ represents the density function of a normal distribution with average μ_2 and variance σ_2^2 when no gene is being expressed, and p is a parameter representing the mixing ratio. [0009]

In regard to the analysis of the experimental errors, there have been proposed several methods of removing systematic errors, so-called normalization methods. For example, when referring to prior art document 3 (Chen et. al., 1997, *Journal of Biomedical Optics*, Vol. 2, pp. 364-374), Chen et. al. assumes that the median values of gene expression levels of two cells are equal to correct for the measured values obtained from channel 1 and channel 2, respectively. Further, when referring to prior art document 4 (Dudoit et. al., 2000, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," Technical Report #5782), prior art document 5 (Schuchhardt et. al., 2000, *Nucleic Acids Research*, Vol.28, No. 10), and prior art document 6 (Yang et. al., 2002, *Nucleic Acids Research*, Vol.30, No.4), Dudoit, Schuchhardt, and Yang consider that systematic errors are caused by different locations of spots on glass slides or different sensitivities of the two kinds of fluorescence dyes, and propose methods of removing the errors.

[0010] The above-mentioned prior art problems are derived from the fact that the analytical results of microarray data lack reproducibility because of low precision and efficiency. It is considered that the cause is insufficient separation of microarray data into true signals, and systematic and measurement errors in the conventional analytical methods. Therefore, removal of systematic errors and evaluation of measurement errors are important issues.

[0011] In regard to the removal of systematic errors, a copending patent application entitled "Method and System for Correction of cDNA

Microarray Data, and Recoding Medium Therefor" has been filed separately. Therefore, the present invention assumes that systematic errors are already removed from the microarray data.

[0012] The conventional analytical methods using microarray data deal with only the ratio (the difference of logarithmic values) of gene expression levels of two channels, that is, they do not deal with the gene expression level of each channel, for the reason that quantitative uniformity of cDNA in each spot is not ensured. Therefore, the conventional analytical methods results in insufficient separation between true signals related to gene expression state and measurement errors.

Summary of the Invention:

[0013] It is an object of the present invention to provide a method and system for separating true signals related to gene expression from measurement errors to increase the precision and efficiency of analysis using microarray data, and further estimating the probability of gene expression in each channel.

[0014] A gene expression state estimating system of the present invention includes an input device for inputting microarray data, a program-controlled data analyzer, and an output device. The data analyzer has parameter estimating means for estimating distributed parameters for each component of a mixed normal distribution and a mixing ratio parameter using gene expression level data given from the input device, and posterior probability calculating means for calculating the posterior probabilities of gene expression in each channel using each of the estimated parameters. The calculated posterior probabilities are outputted to the output device.

[[0015] The adoption of such a configuration to estimate the state of gene expression can attain the object of the present invention.

Brief Description of the Drawings:

[0016] Fig. 1 is a schematic graph of a mathematical model using S-D plots according to the present invention.

[0017] Fig. 2 is a block diagram showing the structure of a first embodiment according to the present invention.

[0018] Fig. 3 is a flowchart showing the operation of the first embodiment according to the present invention.

[0019] Fig. 4 is a block diagram showing the structure of a second embodiment according to the present invention.

[0020] Fig. 5 is a cumulative distribution graph showing an estimated normal distribution of gene expression level data near V=0.

[0021] Fig. 6 is a graph showing the density function of the estimated normal distribution of the gene expression level data near V=0.

[0022] Fig. 7 is a graph showing S-D plots of gene expression level data.

Description of the Preferred Embodiments:

[0023] A mathematical model of gene expression level data obtained from a microarray according to the present invention will first be described. If X denotes the gene expression level of cell 1 obtained with channel 1 and Y denotes the gene expression level of cell 2 obtained with channel 2, then respective gene expression level data are shown in the following equation (2)

$$X = \tau_1\alpha + \beta + \varepsilon_1$$

$$Y = \tau_2\alpha + \beta + \varepsilon_2 \qquad (2)$$

where X and Y denote amounts subjected to adequate transformations of observed values including logarithmic transformation or power transformation and linear transformation.

[0024] Here, $\tau 1$ and $\tau 2$ take either 1 or 0, which represents the presence or absence (ON/OFF) of true gene expression in each cell. Further,

α denotes the amount of mRNA produced when the gene is ON-state and a random variable of gene expression defined by the state of the spot, β denotes a common measurement error between the channel 1 and the channel 2, and ε denotes a measurement error independent between the channels. Note that each distribution of random variables follows the following equation (3)

$$\begin{aligned} &\log \alpha \sim N\left(\mu - \frac{\lambda^2}{2}, \lambda^2\right) \\ &\varepsilon_j \sim N\left(0, \sigma_\varepsilon^2\right), \; j = 1, 2 \\ &\beta \sim N\left(0, \sigma_\beta^2\right) \end{aligned} \qquad (3)$$

[0025] Here, $N(\mu, \sigma^2)$ demotes a one-dimensional normal distribution with average μ and variance σ^2. Further, α, β, and ε are all independent. In this mathematical model, when a gene is being expressed (ON-state), the true expression level is a random variable that takes on nonnegative values, while when it is not being expressed (OFF-state), only simple measurement errors are considered to be observed. Further, referring to the prior art document 6, an S-D transformation is performed as a modification from the M-A transformation adopted by Yang Y.H. et. al. as shown in the following equation (4):

$$\begin{aligned} U &= X + Y, \\ V &= X - Y \end{aligned} \qquad (4)$$

[0026] In other words, the transformation is made assuming that U and V are the sum and difference of gene expression levels of two channels, respectively. A schematic graph of this S-D transformation model is shown in Fig. 1. Note that this plot is called the S-D plot. In Fig. 1, g_{00} represents a simultaneous distribution when no gene is being expressed in both cells, g_{10} represents a simultaneous distribution when a gene is being expressed in cell 1 but not in cell 2, g_{01} represents a simultaneous distribution when a gene is being expressed in cell 2 but not in cell 1, and g_{11} represents a simultaneous

distribution when any gene are being expressed in both cells. The density function of the distribution g_{00} is shown in the following equation (5)

$$g_{00}(u, v \mid \theta) = \varphi\left(u \mid 4\sigma_\beta^2 + 2\sigma_\varepsilon^2\right)\varphi\left(v \mid 2\sigma_\varepsilon^2\right) \quad (5)$$

[0027] Here, $\phi(u|\sigma^2)$ is the density function of a one-dimensional normal distribution with average 0 and variance σ^2. The density function of the distribution g_{10} is shown in the following equation (6):

$$g_{10}(u, v \mid \theta)$$
$$= \int_{-\infty}^{+\infty} \varphi\left(u - \mu e^{-\frac{\lambda^2}{2}+\lambda z} \mid 4\sigma_\beta^2 + 2\sigma_\varepsilon^2\right)\varphi\left(v - \mu e^{-\frac{\lambda^2}{2}+\lambda z} \mid 2\sigma\frac{2}{\varepsilon}\right)\varphi(z \mid 1)dz$$
$$\cong \varphi_2(u - \mu, v - \mu \mid \Sigma_{10}) \quad (6)$$

[0028] Here, $\phi^2(u, v|\Sigma)$ is the density function of a two-dimensional normal distribution with average vector 0 and variance-covariance matrix Σ, andΣ_{10} is a 2×2 variance-covariance matrix, which is shown in the following equation (7)

$$\Sigma_{10} = \begin{pmatrix} \mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2 & \mu^2(e^{\lambda^2} - 1) \\ \mu^2(e^{\lambda^2} - 1) & \mu^2(e^{\lambda^2} - 1) + 2\sigma_\varepsilon^2 \end{pmatrix} \quad (7)$$

[0029] The density function of the distribution g_{01} is shown in the following equation (8):

$$g_{01}(u, v \mid \theta)$$
$$= \int_{-\infty}^{+\infty} \varphi\left(u - \mu e^{-\frac{\lambda^2}{2}+\lambda z} \mid 4\sigma_\beta^2 + 2\sigma_\varepsilon^2\right)\varphi\left(v - \mu e^{-\frac{\lambda^2}{2}+\lambda z} \mid 2\sigma\frac{2}{\varepsilon}\right)\varphi(z \mid 1)dz \quad (8)$$
$$\cong \varphi_2(u - \mu, v - \mu \mid \Sigma_{01}) \quad (8)$$

[0030] Here, Σ_{01} is a 2×2 variance-covariance matrix, which is shown in the following equation (9):

$$\Sigma_{01} = \begin{pmatrix} \mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2 & -\mu^2(e^{\lambda^2} - 1) \\ -\mu^2(e^{\lambda^2} - 1) & \mu^2(e^{\lambda^2} - 1) + 2\sigma_\varepsilon^2 \end{pmatrix} \quad (9)$$

[0031] The density function of the distribution g_{11} is shown in the following equation (10)

$$g_{11}(u, v \mid \theta) = \varphi(v \mid 2\sigma_\varepsilon^2) \int_{-\infty}^{+\infty} \varphi\left(u - \mu e^{-\frac{\lambda^2}{2} + \lambda z} \mid 4\sigma_\beta^2 + 2\sigma_\varepsilon^2\right) \varphi(z \mid 1)\, dz$$

$$\cong \varphi(u - 2\mu \mid 4\mu^2(e^{\lambda^2} - 1) + 4\sigma_\beta^2 + 2\sigma_\varepsilon^2)\, \varphi(v \mid 2\sigma_\varepsilon^2) \quad (10)$$

[0032] Based on the above-mentioned distributions, posterior probabilities of gene expression in cell 1 and cell 2 are shown in the following equations (11) and (12)

$$Pr(\tau_1 = 1 \mid p, \theta) = \frac{p_{10}g_{10}(u, v \mid \theta) + p_{11}g_{11}(u, v \mid \theta)}{f(u, v \mid p, \theta)} \quad (11)$$

$$Pr(\tau_2 = 1 \mid p, \theta) = \frac{p_{01}g_{01}(u, v \mid \theta) + p_{11}g_{11}(u, v \mid \theta)}{f(u, v \mid p, \theta)} \quad (12)$$

where f(u, v|p, θ) is given by the following equation (13)

$$f(u, v \mid p, \theta) = \sum_{(j,k) \in \{0,1\}^2} p_{jk}g_{jk}(u, v \mid \theta) \quad (13)$$

[0033] Note that $p=(p_{00}, p_{10}, p_{01}, p_{11})$ is a parameter representing the mixing ratio for each distribution.

[0034] An embodiment of the present invention will next be described in detail with reference to the accompanying drawings. Referring to Fig. 2, the first embodiment of the present invention is a system for estimating posterior probabilities of gene expression states in cell 1 and cell 2 through the process of formulating a mathematical model related to gene expression level data and the process of estimating unknown parameters by the application of the

formulated mathematical model to the data analysis, and using the calculated estimates of parameters. The system includes an input device 1 such as a keyboard, a program-controlled data analyzer 2, and an output device 3 such as a display device or printer.

[0035] The data analyzer 2 is provided with distributed parameter estimating means 21, mixing ratio parameter estimating means 22, and posterior probability calculating means 23. The distributed parameter estimating means 21 estimates distributed parameters for each component in a mixed normal distribution using gene expression level data from the input device 1. The estimated distributed parameters are sent to the mixing ratio parameter estimating means 22 and the posterior probability calculating means 23. The mixing ratio parameter estimating means 22 estimates a mixing ratio parameter for the mixed normal distribution by a conditional maximum likelihood method using the gene expression level data from the input device 1 and the distributed parameters for each component given from the distributed parameter estimating means 21. The estimated mixing ratio parameter is sent to the posterior probability calculating means 23. The posterior probability calculating means 23 calculates the posterior probability of a gene expression state in each channel using the gene expression level data from the input device 1, the distributed parameters for each component given from the distributed parameter estimating means 21, and the mixing ratio parameter from the mixing ratio parameter estimating means 22. The calculated posterior probability is sent to the output device 3.

[0036] Referring next to Figs. 2 and 3, the process of formulating a mathematical model related to gene expression level data and the process of estimating unknown parameters by the application of the formulated mathematical model to the data analysis will be described in detail. Gene expression level data $\{ (u_i, v_i) \mid i = 1, \ldots, n\}$ given from the input device 1 is sent to the distributed parameter estimating means 21 and the mixing ratio

parameter estimating means 22. The distributed parameter estimating means 21 estimates ξ, μ_0, σ_0, μ_1, and σ_1 by applying the mixed normal distribution of two components, as shown in the following equation (14), to data $\{u_i \mid |v_i| \leq c_M,\ i=1,\ldots,n\}$ on the sum of the amounts of expression of genes near V=0 where c_M denotes the median value of the absolute difference $|v_i|$ $(i = 1,\ldots,n)$ of gene expression levels (step A1 in Fig. 3)

$$(1-\xi)\,\varphi(u-\mu_0 \mid \sigma_0^2) + \xi\varphi(u-\mu_1 \mid \sigma_1^2) \qquad (14)$$

where $\varphi(* \mid \sigma^2)$ is the density function of a one-dimensional normal distribution with average 0 and variance σ^2, (μ_0, σ_0^2) and (μ_1, σ_1^2) are average and variance parameters for first and second components, respectively, and ξ is the mixing ratio, with the assumption that $\mu_0 < \mu_1,\ \sigma_0^2 > 0,\ \sigma_1^2 > 0,\ 0 < \xi < 1$ is satisfied.

[0037] Next, the distributed parameter estimating means 21 uses the estimated $\hat{\xi}, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\mu}_1, \hat{\sigma}_1^2$ to estimate $\mu, \sigma_\varepsilon^2, \sigma_\beta^2, \lambda$ according to the following equations (15), (16), (17), and (18) (step A2)

$$\hat{\mu} = (\hat{\mu}_1 - \hat{\mu}_0)/2 \qquad (15)$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{2\|N_0\|}\sum_{i\in N_0} v_i^2 \qquad (16)$$

$$\hat{\sigma}_\beta^2 = \frac{1}{4}\hat{\sigma}_0^2 - \frac{1}{2}\hat{\sigma}_\varepsilon^2 \qquad (17)$$

$$\hat{\lambda} = \sqrt{\log\left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{4\hat{\mu}^2}\right)} \qquad (18)$$

where N_0 denotes an index set of data values that satisfies $i \in \{i \mid u_i < \hat{\mu}_0\}$ and $\|N_0\|$ denotes the number of elements.

[0038] Next, the mixing ratio parameter estimating means 22 estimates a mixing ratio parameter $p=(p_{00}, p_{10}, p_{01}, p_{11})$ by a conditional maximum likelihood method using an estimate $\hat{\theta} = (\hat{\mu}, \hat{\lambda}, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_\beta^2)$ of each parameter given from the distributed parameter estimating means 21 by applying a

two-variable mixed normal distribution of four components shown in the following equation (19) to the gene expression level data $\{(u_i, v_i) \mid i=1,\ldots,n\}$ given from the input device 1 (step A3).

$$\begin{aligned} & p_{00}g_{00}(u, v \mid \hat{\theta}) + p_{10}g_{10}(u, v \mid \hat{\theta}) + p_{01}g_{01}(u, v \mid \hat{\theta}) + p_{11}g_{11}(u, v \mid \hat{\theta}) \\ & = p_{00}\varphi(u \mid 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2)\varphi(v \mid 2\hat{\sigma}_\varepsilon^2) + p_{10}\varphi_2(u - \hat{\mu}, v - \hat{\mu} \mid \Sigma_{10}) \\ & + p_{01}\varphi_2(u - \hat{\mu}, v + \hat{\mu} \mid \Sigma_{01}) + p_{11}\varphi(u - 2\hat{\mu} \mid 4\hat{\mu}^2(e^{\lambda^2} - 1) \\ & + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2)\varphi(v \mid 2\hat{\sigma}_\varepsilon^2) \end{aligned} \tag{19}$$

[0039] Here, it is assumed that the above equation satisfies the relationships shown in the following equation (20) (where $\hat{\Sigma}_{10}$ is a 2x2 variance-covariance matrix derived from the equation (7)) and the following equation (21) (where $\hat{\Sigma}_{01}$ is a 2x2 variance-covariance matrix derived from the equation (9)).

$$\hat{\Sigma}_{10} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2 & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_\varepsilon^2 \end{pmatrix} \tag{20}$$

$$\hat{\Sigma}_{01} = \begin{pmatrix} \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 4\hat{\sigma}_\beta^2 + 2\hat{\sigma}_\varepsilon^2 & -\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) \\ -\hat{\mu}^2(e^{\hat{\lambda}^2} - 1) & \hat{\mu}^2(e^{\hat{\lambda}^2} - 1) + 2\hat{\sigma}_\varepsilon^2 \end{pmatrix} \tag{21}$$

[0040] The process of estimating posterior probabilities of gene expression states in cell 1 and cell 2 using the calculated estimates of parameters will next be described.

[0041] The posterior probability calculating means 23 can describe the posterior probabilities of gene expression state in each cell for each pair (u, v) of the gene expression level data given from the input device 1 using the estimates $\hat{\theta} = (\hat{\mu}, \hat{\lambda}, \hat{\sigma}_\varepsilon^2, \hat{\sigma}_\beta^2)$ and $\hat{p} = (\hat{p}_{00}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{11})$ of each parameter given from the distributed parameter estimating means 21 and the mixing ratio parameter estimating means 22.

[0042] In other words, the posterior probabilities indicating that any gene expression is ON-state in cell 1 and cell 2 can be calculated from the

following equations (22) and (23) (step A4).

$$\Pr(\tau_1 = 1 \mid \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10} g_{10}(u, v \mid \hat{\theta}) + \hat{p}_{11} g_{11}(u, v \mid \hat{\theta})}{f(u, v \mid \hat{p}, \hat{\theta})} \quad (22)$$

$$\Pr(\tau_2 = 1 \mid \hat{p}, \hat{\theta}) = \frac{\hat{p}_{01} g_{01}(u, v \mid \hat{\theta}) + \hat{p}_{11} g_{11}(u, v \mid \hat{\theta})}{f(u, v \mid \hat{p}, \hat{\theta})} \quad (23)$$

[0043] It is then judged whether calculations of posterior probabilities indicating that any gene expression is ON-state have been made for all the pairs (u, v) of the gene expression level data (step A5). When all the calculations have been completed, the process is ended, while when all the calculations have not been completed yet, the posterior probability related to the next gene is calculated.

[0044] The calculated posterior probabilities of gene expression in each channel are sent to the output device 3. The output device 3 displays or prints out the posterior probabilities of gene expression in each channel in the form of a graph.

[0045] The following describes the effects of the embodiment. In the embodiment, a mathematical model in which the concept of gene expression/nonexpression is introduced is constructed to separate true signals from experimental errors. Further, the use of data on the sum and difference of gene expression levels in two channels makes it easy to obtain information on the sensitivity of microarray data to fluorescence intensities in each channel, allowing more accurate extraction of the magnitude of experimental errors. Furthermore, a two-dimensional simultaneous distribution is described for these sum and difference data. It allows high-precision estimation of posterior probabilities of gene expression in each channel.

[0046] In addition, the posterior probability indicating an event of differential expression between cell 1 and cell 2 (mismatched ON-OFF state) (step 4 in Fig. 3) is calculated by the following equation (24).

$$\Pr(\tau_1 \neq \tau_2 \mid \hat{p}, \hat{\theta}) = \frac{\hat{p}_{10} g_{10}(u, v \mid \hat{\theta}) + \hat{p}_{01} g_{01}(u, v \mid \hat{\theta})}{f(u, v \mid \hat{p}, \hat{\theta})} \quad (24)$$

[0047] Thus, the embodiment has the advantage of detecting candidate genes that are likely to reveal differential expression in cell 1 and cell 2.

[0048] A second embodiment of the present invention will next be described in detail with reference to Fig. 4. Like the first embodiment, the second embodiment of the present invention includes the input device, data analyzer, and the output device. The second embodiment also includes a recording medium 4 with a data analyzing program recorded on it. The recording medium 4 may be either portable or fixed type, such as a magnetic disk, semiconductor memory, CD-ROM, or any other recording medium. Alternatively, a computer program capable of executing the method of the present invention may be stored in a memory device of a computer connected to a network so that it can be transferred to another computer through the network. The form of the medium that provides a computer program executing the algorithm is a distributable as a medium readable in a variety of computer formats, and is not limited to a specific type.

[0049] The data analyzing program is read from the recording medium 4 into a data analyzer 5 to control the operation of the data analyzer 5 to execute processing on data files inputted from the input device 1 in the same manner as the data analyzer 2 does in the first embodiment.

[0050] The following specifically describes the embodiments of the present invention. Data used as an example is obtained from an experiment for comparing the states of gene expression of two different types of cancer cells (cell 1 and cell 2).

[0051] The test is conducted on expression patterns of 48 grids on one chip, 441 (21×21) spots per grid, a total of 21168 genes.

[0052] Figs. 5 and 6 show estimation results of each of the distributions

U of the sum of expression levels when cell 1 and cell 2 both show OFF or ON state of gene expression (V=0), a mixed normal distribution, and a single-peaked normal distribution for contrast purposes. The following table 1 shows estimation results of distributed parameters for each component (results of step A1).

[Table 1]

Result of N2MIXFit (Ver 0.998)

Name of Data Set to be analyzed = n145h1.std
Name of Target Variate = S_CH1
Type of Transformation = 1/16
Sample size = 14726
Critical Value for Convergence = .1000015-06
Iterations for Convergence = 50
Job Termination Status = Normally Terminated

	Mean	SD	Rate(%)
Single Component:	3.0273	4.3560	100.00
1st Component:	.89956	1.2978	47.38
2nd Component:	4.9430	5.1395	52.62

Log_Likelihood for Single Component = -42565.
for Two Components Mixed = -40465.
Log of Likelihood Ratio Statistics = 2099.3

[0053] Fig. 5 shows cumulative distribution functions and Fig. 6 shows density functions, in which the thin solid line indicates the estimation results of an assumed mixed normal distribution, the chain double-dashed line indicates the estimation results of its first component (OFF-OFF), the bold solid line indicates the estimation results of its second component (ON-ON), and the dashed line indicates the estimation results of an assumed single-peaked normal distribution.

[0054] The long and short dashed line in Fig. 5 indicates an empirical cumulative distribution function based on observed data, showing that it well follows the curve of the mixed normal distribution (thin solid line) in which observed values are estimated.

[0055] In Fig. 6, the asterisk marks (which are replaced with the following hatching patterns (1) to (5) in the range of gene expression levels from

about 0 to 30 because, though the asterisk marks can be discernible at both ends, they are densely overlapped within the range) indicates observed data. The hatching patterns ((1) to (5)) represent the magnitude of posterior probability values belonging to the first component. In other words, the solidly shaded area (1) indicates the range of posterior probabilities from 0 to 0.2, the hatching area (2) from 0.2 to 0.4, the hatching area (3) from 0.4 to 0.6, the hatching area (4) from 0.6 to 0.8, and the hatching area (5) from 0.8 to 1.0.

[0056] Fig. 7 shows an S-D plot of gene expression level data, in which the abscissa indicates the sum of logarithmic values of gene expression levels of cell 1 and cell 2, and the ordinate indicates the different between the logarithmic values. In Fig. 7, the hatching patterns represent the magnitude of posterior probabilities indicating mismatched expression states (ON-OFF or OFF-ON) between cell 1 and cell 2. In other words, the solidly shaded area (1) indicates the range of posterior probabilities from 0 to 0.2, the hatching area (2) from 0.2 to 0.4, the hatching area (3) from 0.4 to 0.6, the hatching area (4) from 0.6 to 0.8, and the hatching area (5) from 0.8 to 1.0.

[0057] The following table (2) shows estimation results of distributed parameters by the conditional maximum likelihood method (results of steps A2 and A3 in Fig. 3). A gene corresponds to each plotted spot in Fig. 7, and this makes it easy to narrow down gene candidates related to the difference between cell 1 and cell 2.

[Table 2]

RESULT OF MAD Var(0.998)

Transformation = 6 (1/16)
Sample Size = 21168

SIGMA_Epsilon =	.838
MU =	2.022
LAMBDA =	.950
SIGMA_Beta =	.133
P11 =	.561
P10 =	.004
P01 =	.011
P00 =	.425

[0058] The first effect of the present invention is to achieve a separation between true signals related to gene expression and experimental errors by introducing the concept of gene expression/nonexpression into the gene expression level data obtained from microarrays to construct a mathematical model.

[0059] The second effect of the present invention is to make it easy to obtain sensitivity information on the sensitivity of microarray data to fluorescence intensities of two channels by transforming the gene expression level data obtained from microarrays into data on the sum and difference of the gene expression levels between two channels. It then makes it possible to visualize the magnitude of experimental errors.

[0060] The third effect of the present invention is to enable estimation of posterior probability related to expression/nonexpression of each gene in each of the two channels by transforming the gene expression level data obtained from microarrays into data on the sum and difference of the gene expression levels between two channels to describe a two-dimensional simultaneous distribution of the sum and difference data. It then makes it possible to detect genes related to differences between cell 1 and cell 2 with high precision.